# Energy-Efficient Logic and SRAM Design: a Case Study

Nele Reynders, Bram Rooseleer and Wim Dehaene
*KU Leuven, ESAT-MICAS*
*Email: nele.reynders@esat.kuleuven.be*

***Abstract*—This paper discusses energy-efficient design, both for logic and for memories. For datapaths which are controlled by dynamic energy, high energy-efficiency can be obtained by significantly reducing the supply voltage. However, simply lowering $V_{\mathrm{DD}}$ does not automatically imply more energy-efficient operation for SRAM memories, as they are dominated by static leakage. This paper identifies which design methodologies can be employed to achieve high energy-efficiency. In particular, a JPEG encoder fabricated in a $40\,\mathrm{nm}$ CMOS technology is used as a case study to determine the trade-offs, challenges and benefits of energy-efficient design.**

## I. Introduction

The energy consumption of digital circuits is of crucial importance for many applications. In order to design energy-efficient circuits, it is imperative to determine which component of energy is dominant for which type of circuit or system. The energy consumption of digital circuits can be divided into two components, i.e. dynamic and static energy:

$$E_{\mathrm{tot}} = E_{\mathrm{dyn}} + E_{\mathrm{stat}} \quad (1)$$

$$= \alpha \cdot C \cdot V_{\mathrm{DD}}^2 + I_{\mathrm{leak}} \cdot V_{\mathrm{DD}} \cdot t_{clk} \quad (2)$$

where $\alpha$ is the circuit activity and $t_{clk}$ the clock period. The dynamic energy consists mostly of switching energy when load capacitances (depicted by $C$) are charged or discharged. Static energy, on the other hand, is consumed as a result of all sources of leakage (combined in $I_{\mathrm{leak}}$) during a certain time period. The dynamic energy is quadratically dependent on the supply voltage. Leakage power also decreases with reducing $V_{\mathrm{DD}}$ but at the same time the clock period increases. By lowering $V_{\mathrm{DD}}$, $E_{\mathrm{dyn}}$ decreases significantly, but $E_{\mathrm{stat}}$ will increase due to the larger $t_{clk}$. Therefore, there exists an optimal supply point at which $E_{\mathrm{tot}}$ is minimal. Depending on the activity of a circuit, this optimum occurs at lower or higher $V_{\mathrm{DD}}$.

Hence, extremely reducing the supply voltage is mostly interesting for circuits which are dominated by dynamic energy consumption. E.g., high performance datapaths exhibit large dynamic energy dominance. However, the feasible performance of such circuits decreases significantly with $V_{\mathrm{DD}}$. Therefore, possible applications for ultra-low-voltage operation are very energy-constrained systems which exhibit less severe performance constraints. However, the performance reduction by lowering the supply can be mitigated by decreasing the threshold voltage as well. This is the reason why sub- and near-threshold circuits with very low energy consumption that still achieve a moderately high performance use low-$V_{\mathrm{T}}$ (LVT) transistors and decrease their $V_{\mathrm{DD}}$ considerably [1], [2].

Memories on the other hand are dominated by leakage, since only few cells are accessed simultaneously. Decreasing dynamic energy by reducing the supply voltage thus only has a limited effect, while the increased delay has a detrimental impact on the leakage energy. In addition, reducing the threshold voltage to reach sufficient speed aggravates this problem.

To summarize, energy-efficient design is not simply equal to ultra-low supply voltage design. Depending on the application at hand and on its activity factor, different measures should be taken to perform a given operation with minimal energy while still meeting the performance requirements. An interesting analysis on this subject can also be found in [3]. This paper extensively covers the various trade-offs which emerge in energy-efficient designs. The case study used for this discussion is a near-threshold JPEG encoder [4] processed in $40\,\mathrm{nm}$ CMOS. Section II focuses on the design of energy-efficient logic, while Section III discusses energy-efficient memory implementations. Section IV concludes the paper.

## II. Ultra-Low-Voltage Logic

High energy-efficiency can be achieved by operating a circuit at or around its minimum energy point (MEP), which typically occurs at supply voltages much lower than the nominal $V_{\mathrm{DD},nom}$. However, as explained in Section I, the circuit's delay decreases significantly when lowering $V_{\mathrm{DD}}$. Therefore, only applications which are less stringent on performance than they are on energy requirements are suitable for sub- or near-threshold design. Fortunately, many applications exist that do not need a speed of hundreds of $\mathrm{MHz}$ or more. For example, in applications like RFID tags, biomedical devices and sensor nodes, energy-efficiency is more crucial. Nonetheless, these applications often require moderate performance in the order of tens of $\mathrm{MHz}$ and are not satisfied with only $\mathrm{kHz}$ speed. This paper therefore focuses on combining ultra-low energy consumption with moderately high performance.

A very important challenge for ultra-low-voltage designs is the sensitivity to variations of such circuits. The increased sensitivity to variations compared to circuits operating at $V_{\mathrm{DD},nom}$ can severely compromise the robustness of circuits and the overall yield. As a result, it is imperative to design variation-resilient circuits.

### A. Case study: JPEG encoder

The JPEG encoder used as case study consists of several building blocks (Fig. 1): 2-Dimensional Discrete Cosine Trans-

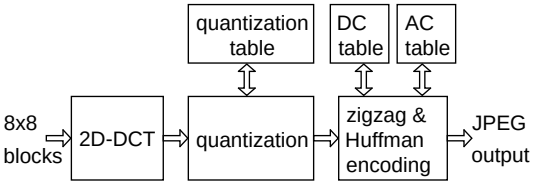


Fig. 1. Block diagram of the JPEG encoder.

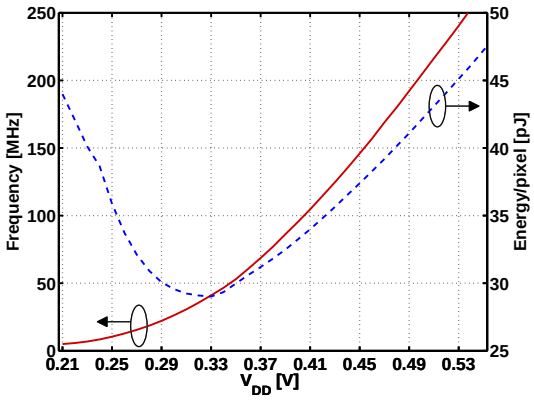


Fig. 2. Measured maximum operating frequency and energy consumption per pixel as function of $V_{\mathrm{DD}}$. This concerns mean values, obtained from measuring 26 dies.

form (2D-DCT), a quantization block, a zigzag block which performs reordering of the transformed and quantized 8x8 input blocks and Huffman encoding to efficiently encode the zigzagged coefficients. Essential to know is whether a building block contains simply a logic function or if it also needs memory. The 2D-DCT can be seen as a large and complex datapath, while both the quantization and the zigzag & Huffman have relatively less logic but require stored data in lookup tables (LUT). The quantization divides transformed input blocks by quantization coefficients which are stored in a quantization table. The zigzag & Huffman block fetches the correct codes saved in the DC and AC Huffman tables.

We will first discuss the design style of the datapath components which perform arithmetical or logical operations. In the JPEG encoder, a latch-based pipelined architecture is used, combined with transmission gate (TG) logic for the logic gates. TG logic is employed because it exhibits several attractive properties for ultra-low-voltage operation [5]. The most beneficial one is its high variation-resilience due to the inclusion of both pMOS and nMOS transistors in each conducting path. Other interesting properties are low contribution to leakage and low area occupation due to almost minimally sized transistors, as opposed to standard CMOS logic. The TG logic is implemented differentially to be able to cascade multiple logic gates because they require differential input signals. By cascading logic gates, averaging of timing variations is obtained, thus reducing timing variability. The differential character adds significantly to the overall variation-resilience of the system.

TG logic exhibits some signal loss at its output, therefore such logic gates cannot be infinitely cascaded without regeneration. This regeneration of the signal swing can be performed for instance by 2 inverters on the complementary outputs or by a latch. The overhead of the latches is very limited in this architecture, since they are efficiently implemented with the same 2 inverters and a few extra transmission gates. These latches do not only solve the need for regeneration, but they also create a pipelined architecture. Pipelining increases the throughput of a system, which is valuable in ultra-low-voltage designs because of their inherently low speed. Of course, flipflop-based pipelining would be an option as well, but latches allow time borrowing. Seeing the fact that timing variations of sub- and near-threshold designs are very high, time borrowing is a very beneficial concept. Unfortunately, timing verification tools do not yet standardly support latches.

The employed architecture is deeply pipelined. By using short pipeline stages, not only the throughput is enhanced, but the leakage energy also reduces significantly due to the smaller clock period. This is even more pronounced in ultra-low-voltage designs because of the inherently high circuit delay. Thereby, aggressive pipelining in sub- or near-threshold designs effectively shifts the MEP to a lower supply voltage and to a lower absolute value of the total energy consumption [6]. To summarize, the combination of TG logic and latch-based pipelining is employed in the near-threshold JPEG encoder for optimal throughput and variation-resilience.

The memory elements are implemented as register matrices in the fabricated version of the JPEG encoder, because the speed of sub- or near-threshold SRAM is not high enough. Large area and leakage reductions could be achieved by allowing an SRAM to operate at a $2^{nd}$, higher supply (see Section III). The sizes of the 3 LUTs used in the JPEG encoder are 64 words of 20 bit for the quantization table, 12 words of 20 bit for the DC Huffman table and 162 words of 20 bit for the AC Huffman table. A table consists of a decoder, a register matrix and word selectors. The decoders receive an address from the datapath and drive the word lines for the word selectors. The register matrix is serially written at startup and is from then on only accessed for reading words. A single table register consists of a master and a slave latch. Since the master latch is only necessary during writing at startup, it is power gated afterwards, while the slave latch stores the data. The power gating of the master latch aids to significantly reduce the leakage of the register matrix. Both the decoder and the word selector are implemented using the same design style as the rest of the JPEG encoder.

### *B. Case study: measurement results*

Measurements of the JPEG encoder which was fabricated in a 40 nm CMOS technology show that the chip is fully functional down to a supply voltage of 210 mV. The targeted speed in the range of tens of MHz has been obtained, as visible in Fig. 2. At the minimal supply, a frequency of 5 MHz is achieved. Frequencies of up to 275 MHz are possible with

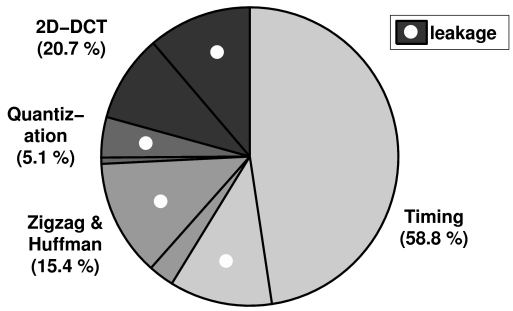


Fig. 3. Energy division at the MEP ($V_{\mathrm{DD}} = 330\,\mathrm{mV}$). The contribution of leakage to the different building blocks is also indicated.

supplies from 210 to 550 mV.

The aim of an energy-efficient design has also been realized, as can be seen in Fig. 2 where the energy consumption per pixel is visualized. The energy consumption is below 50 pJ per pixel for supplies of up to 550 mV. The MEP of 29.01 pJ per pixel occurs at a 330 mV supply and a 41 MHz speed. Naturally, the highest energy-efficiency is obtained at the MEP, but interesting is that the region around the MEP is quite flat. For supplies from 290 to 350 mV, the energy consumption stays below 30 pJ. Depending on the desired operating frequency, a high energy-efficiency can thus still be obtained in a relatively large region around the MEP.

Measurements on the JPEG encoder were performed on 26 dies, in order to be able to adequately study the variations. Fig. 2 shows the obtained mean results. The variation-resilience of the chip is verified through the measurements: for the $210-550\,\mathrm{mV}$ supply range, the variation $\sigma/\mu$ in operating frequency and energy consumption per pixel is only $8.6\%$ and $5.4\%$, respectively.

Fig. 3 shows the contribution of the different building blocks to the total energy consumption at the MEP. The timing block, which consists of a non-overlapping clock generator and a clock tree, consumes $58.8\%$ of the energy. The 3 other building blocks combined consume less than half of the total energy. The contribution of leakage to each block's energy consumption can also be seen in Fig. 3. Although design effort was made to reduce the leakage of the tables as much as possible, it is still apparent that the quantization and zigzag & Huffman blocks have a much higher percentage of leakage than the 2D-DCT (which did not have a table) does.

## III. Energy-Efficient SRAM

In Fig. 3, it can be seen that a large portion of the used energy is caused not by active switching but rather by leakage. The quantization and the Huffman encoder blocks both have a share of leakage energy of 60 % or more, even at the highest speed. These are the two blocks that contain lookup tables. These tables have, due to their nature, a very low activity. Nevertheless, they are designed using the same methodology as the rest of the JPEG encoder, resulting in very large register banks. Reducing the leakage energy of these circuits is thus very important to reach an energy-efficient design.

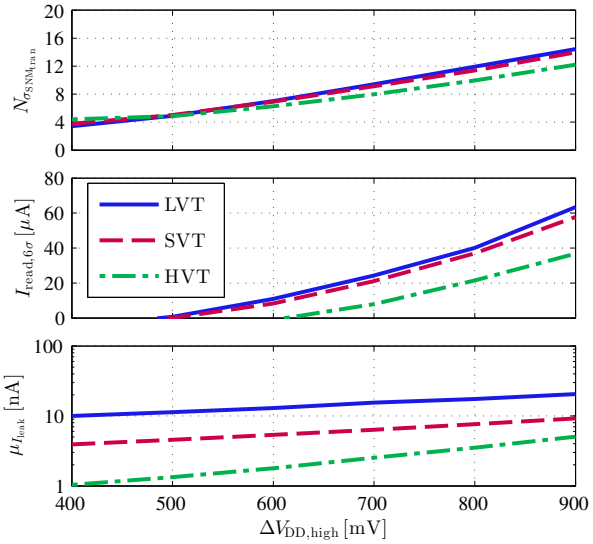


Fig. 4. 6T SRAM cell parameters as function of supply and threshold voltage.

### A. Low energy SRAM design

In many other applications, this leakage problem is even much higher as the size of their memories are much larger. The solution lies in revisiting the reasoning that led to the adoption of near-threshold design. In this reasoning, high activities are assumed. Therefore, reducing active energy has been given priority. This reduction is achieved by decreasing the supply voltage as much as possible. To reach sufficient speed, LVT transistors are used. However, in the case of memories, not active energy but leakage is dominant. Reducing $V_{\mathrm{DD}}$ reduces leakage, but only to a much smaller extend than $V_{\mathrm{T}}$. The best way of reducing leakage is by using high threshold voltages (HVT) while increasing $V_{\mathrm{DD}}$ to keep speed constant. Fig. 4 shows the mean leakage, the worst-case read current and the stability of a classic 6T-SRAM cell as function of the supply voltage for three different threshold voltages. It is clear that for the same read current (and thus speed), increasing both $V_{\mathrm{T}}$ and $V_{\mathrm{DD}}$ is beneficial both for leakage and for stability [7].

Aside from combining higher supply voltages with HVT transistors for the memory cells, other techniques are used to reduce the active energy or leakage power of a memory. Here a short overview of possible circuit techniques is given.

Word lines can be divided into local word lines [8], that only activate the word to be read or written. This highly reduces wasteful activity. Most active energy in a memory is used by the data transfers on the bit lines. These bit lines are highly capacitive due to the large number of connected cells. Dividing these bit lines in local bit lines connected to a single global bit line reduces this load significantly [9]. The energy use on these lines can be further reduced by using low swing signals. When using the memory in a near-threshold design, the supply of the near-threshold circuit might be reused for this. Fig. 5 shows different possibilities of dividing bit lines [10]. Method (a) connects them with a simple pass transistor. This is a very

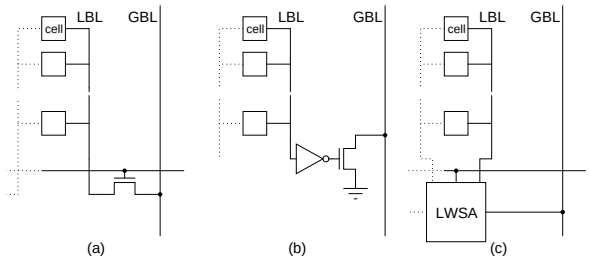


Fig. 5. Different ways of connecting local bit lines (LBL) and global bit lines (GBL).

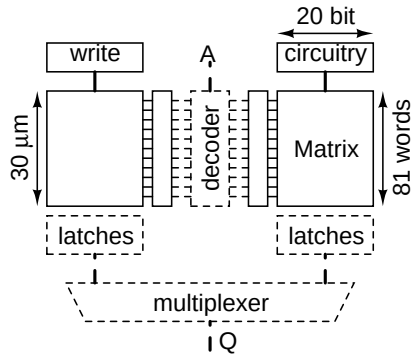


Fig. 6. Proposed architecture of a LUT.

simple method, but it requires the same swing on the global and the local bit lines. Method (b) can only be used for the read operation. A full swing signal is used on the (small) local bit lines and a low swing signal on the large global bit lines. The memory cell does not have to drive the large global bit line anymore, speeding up operation as well. The third method (c) uses a local write sense amplifier (LWSA) to be able to use low swing global bit lines for write operation. Normally, the last 2 options are combined. The disadvantage of the combination of these methods is a relatively large increase in area.

### *B. Case study: LUTs for the JPEG encoder*

As a case study, the design of the AC Huffman table with $162$ words of $20$ bit is considered. With only $3240$ bits, this is a very small memory. 6T SRAM cells with HVT transistors can be used with a matrix voltage $V_{\mathrm{DD,\,high}}$ of $700\,\mathrm{mV}$. This results in a read current sufficiently high to reach the $41\,\mathrm{MHz}$ MEP speed and sufficient stability (see Fig. 4). To shorten the bit lines, each row contains two words. The decoder is placed in the middle of the matrix as shown in Fig. 6. Bit lines can be divided in $10$ blocks of $8$ words and a single extra word. As these tables are not written much, bit lines can be divided using method (a) of Fig. 5. The $330\,\mathrm{mV}$ $V_{\mathrm{DD}}$ supply of the JPEG encoder is used as a low swing voltage for read operation. Writing is done with full swing signaling, but this has no influence on energy as it is only done during startup.

The most active part, i.e. the decoder, has $V_{\mathrm{DD}}$ as supply and is implemented using the same design style as the rest of the JPEG encoder. Level shifters are used to shift the output of the decoder to $V_{\mathrm{DD,\,high}}$. As $V_{\mathrm{DD}}$ is used as the low swing voltage on the bit lines, these have the same logic levels as the JPEG encoder. No sense amplifiers or level shifters are needed and the output can simply be captured by latches working at $V_{\mathrm{DD}}$.

It is clear that when allowing a second, higher supply voltage in the JPEG encoder, superior memory architectures (such as Fig. 6) can be employed, yielding large reductions in terms of area and leakage. For example, the cell area and the cell leakage power then both decrease significantly from $13.505$ to $0.475\,\mu\mathrm{m}^2$ and from $6.86$ to $1.73\,\mathrm{nW}$, respectively.

## IV. Conclusion

As a general conclusion, it can be stated that for circuits with a high activity, low supply voltages are needed to keep switching energy low. To achieve sufficient speed, LVT transistors are needed. These near-threshold circuits are vulnerable to variations. A design methodology to cope with this problem has been proposed. This story changes when circuits with a very low activity, such as memories, are used. For these circuits, leakage is the dominant factor and high threshold voltages combined with a high supply are preferred. Some techniques to reduce active energy in memories are given. When the activity of circuits is variable, such as is the case with burst mode circuits, power gating should be used.

## References

[1] N. Reynders and W. Dehaene, "A 190mV supply, 10MHz, 90nm CMOS, pipelined sub-threshold adder using variation-resilient circuit techniques," in *IEEE Asian Solid State Circuits Conference (A-SSCC)*, Nov. 2011, pp. 113–116.

[2] ——, "Variation-resilient sub-threshold circuit solutions for ultra-low-power digital signal processors with 10MHz clock frequency," in *IEEE European Solid-State Circuits Conference (ESSCIRC)*, Sep. 2012, pp. 474–477.

[3] D. Bol, R. Ambroise, D. Flandre, and J. Legat, "Analysis and minimization of practical energy in 45nm subthreshold logic circuits," in *IEEE International Conference on Computer Design (ICCD)*, Oct. 2008, pp. 294–300.

[4] N. Reynders and W. Dehaene, "A 210mV 5MHz variation-resilient near-threshold JPEG encoder in 40nm CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2014, pp. 456–457.

[5] ——, "Variation-resilient building blocks for ultra-low-energy sub-threshold design," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 12, pp. 898–902, 2012.

[6] D. Jeon, M. Seok, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A super-pipelined energy efficient subthreshold 240 MS/s FFT core in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 23–34, 2012.

[7] B. Rooseleer, S. Cosemans, and W. Dehaene, "A 65 nm, 850 MHz, 256 kbit, 4.3 pJ/access, ultra low leakage power memory using dynamic cell stability and a dual swing data link," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 7, pp. 1784–1796, 2012.

[8] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano, "A divided word-line structure in the static RAM and its application to a 64k full CMOS RAM," *IEEE Journal of Solid-State Circuits*, vol. 18, no. 5, pp. 479–485, 1983.

[9] A. Karandikar and K. Parhi, "Low power SRAM design using hierarchical divided bit-line approach," in *IEEE International Conference on Computer Design (ICCD)*, Oct. 1998, pp. 82–88.

[10] B. Rooseleer and W. Dehaene, "A 40 nm, 454MHz 114 fJ/bit area-efficient SRAM memory with integrated charge pump," in *IEEE European Solid-State Circuits Conference (ESSCIRC)*, Sep. 2013, pp. 201–204.